Andrew Young School of Policy Studies
Research Paper Series

Working Paper 09-12
October 2009 (Updated May 2010)

Department of Economics
W.J. Usery Workplace Research Group

A Non-Experimental Evaluation of Curriculum Effectiveness in Math

Rachana Bhatt
Georgia State University

Cory Koedel
University of Missouri

Georgia State University | Andrew Young School of Policy Studies

W.J. Usery Workplace Research Group Paper Series

Working Paper 2009-10-2
October 2009 (Updated May 2010)

A Non-Experimental Evaluation of Curriculum Effectiveness in Math

Rachana Bhatt
Georgia State University

Cory Koedel
University of Missouri

This paper can be downloaded at: http://aysps.gsu.edu/usery/Papers.html

Georgia State University | Andrew Young School of Policy Studies

A Non-Experimental Evaluation of Curricular Effectiveness in Math

Rachana Bhatt
Georgia State University

Cory Koedel*
University of Missouri

May 2010

We use non-experimental data from a large panel of schools and districts in Indiana to evaluate the impacts of math curricula on student achievement. Using matching methods, we obtain causal estimates of curriculum effects at just a fraction of what it would cost to produce experimental estimates. Furthermore, external validity concerns that are particularly cogent in experimental curricular evaluations suggest that our non-experimental estimates may be preferred. In the short term, we find large differences in effectiveness across some math curricula. However, as with many other educational inputs, the effects of math curricula do not persist over time. Across curriculum adoption cycles, publishers that produce less effective curricula in one cycle do not lose market share in the next cycle. One explanation for this result is the dearth of information available to administrators about curricular effectiveness.

* We thank Emek Basker, Julie Cullen, Gordon Dahl, Barry Hirsch, Josh Kinsler, David Mandy, Peter Mueser, Rusty Tchernis and many seminar and conference participants for useful comments and suggestions. We also thank Karen Lane and Molly Chamberlin at the Indiana Department of Education for help with data. This work was not funded or influenced by any outside entity.

I. Introduction

Curricular effectiveness has received much attention in the education literature, and justifiably so (see, for example, Slavin and Lake, 2008; National Research Board, 2004). The majority of instructional time and homework assignments are textbook oriented, and a substantial amount of school expenditures are devoted to curriculum purchases. According to a 2002 survey sponsored by the National Education Association and the American Association of Publishers, 80% of teachers use textbooks in the classroom, and over half of students' in-class instructional time involves textbook use (Finn, 2004).[1] In 2006 alone, expenditures on K-12 instructional materials totaled close to $8.1 billion dollars.[2] Different curricula are developed using different theories about how students learn - this results in different content, organization and structure across curricula for the same subject and grade group. Given the central role that curricula play for students and schools, it is of interest to determine the extent to which different curricula differentially affect student achievement.

Hundreds of studies have attempted to evaluate the curricular alternatives facing school administrators. However, much of the literature on curricular effectiveness lacks scientific rigor, raising concerns about the reliability of the findings. For example, in 2007, the What Works Clearinghouse (WWC), which was established by the Institute for Education Sciences (IES) to serve as a filter for education research, evaluated over 200 studies of curricular effectiveness in elementary mathematics and found that over 96 percent of these studies did not meet reasonable quality standards (WWC, 2007).[3] Likely in response to the dearth of reliable evidence in the

[1] Textbooks are just one component of the curricula purchased by schools from publishers. Other aspects include teacher instructional support services and supplementary materials such as student workbooks, flashcards, and solution manuals.

[2] See http://www.aapschool.org/vp_funding.html

[3] The WWC reviews the literature on a variety of topics in education, including the effects of curriculum adoptions, and classifies studies as either (1) meets evidence standards, (2) meets evidence standards with reservations or (3) does not meet evidence standards. Generally speaking, studies in category (1) use randomized controlled trials (RCTs) or quasi-experiments (e.g., regression discontinuity designs). Studies in category (2) may employ non-experimental techniques, but must be deemed by the principal investigator at WWC to have employed appropriate statistical tools such that causal inference is reasonable. Of the 237 studies on elementary math curricula reviewed by the WWC as of July, 2007, just nine were deemed to be of sufficient quality by WWC to be included in categories (1) and (2) (WWC, 2007).

literature, recent research has turned to randomized controlled trials (RCTs) to evaluate curricular effectiveness (see, for example, Agodini et al., 2009; Borman et al., 2008; Resendez and Azin, 2007). RCTs randomly assign curricula across schools (and/or classrooms) and produce causal estimates of curriculum effects that are internally valid – that is, valid within the framework of the experiment. However, a general drawback of RCTs that is particularly cogent in the case of curricular evaluation is that the estimates may not extrapolate well outside of the experiment.

We highlight two concerns with RCTs in the context of curricular evaluation that will potentially limit the external validity of their results.[4] First, RCTs require voluntary participation by both *schools and curriculum publishers.* If the schools that select into the experiment differ from the general population of schools, then Manski's (1996) "experimentation on a subpopulation" concern is relevant, and the experimental results will not necessarily reveal anything about curricular effectiveness at schools not represented in the study.[5] Perhaps more importantly, there is also a selection problem with respect to publishers because publishers are typically actively involved in the experiments. For example, in recent experimental studies by Agodini et al. (2009), Borman et al. (2008) and Resendez and Azin (2007), publishers directly provided teacher training and support services.[6] The active role of publishers in experimental studies means that publishers must agree to participate, and only publishers that expect their curricula to be successful in the setting of the RCT are likely to do so. Overall, the requirements of voluntary school and publisher participation limit the extent to which experimental designs can be used to evaluate the full curricular landscape.

[4] See Heckman and Smith (1995) and Manski (1996) for general discussions about the strengths and weaknesses of experimental research designs.

[5] A common concern in educational experiments is that participating schools may differ in leadership from the average non-participating school. For instance, in Agodini et al's. (2009) study, the authors state "Participating sites are not a representative sample of districts and schools, because interested sites are likely to be unique in ways that make it difficult to select a representative sample. Interested districts were willing to use all four of the study's curricula, allowed the curricula to be randomly assigned to their participating schools, and were willing to have the study team test students and collect other data required by the evaluation."

[6] In fact, we are not aware of any experimental curriculum evaluations where the publishers were not actively involved.

A second threat to the external validity of RCTs is publisher responsiveness to evaluation, commonly referred to as Hawthorne effects. In the general experimental literature, Hawthorne effects refer to the subjects of the experiment. In the case of curricular evaluation, the active role of publishers suggests that in addition to schools and students, they *are* subjects. Given that the experimental evaluations are high-stakes competitions for publishers, there is no reason to expect them to take a "business-as-usual" approach. The potential for publisher Hawthorne effects in RCTs raises concerns about how well their results will extrapolate to lower-stakes environments for publishers.[7]

In addition to these threats to external validity, the costs associated with RCTs limit the amount of information that they can provide. For example, because RCTs are expensive, they generally focus on just one or two curricula evaluated at small numbers of schools and districts.[8] The expenses associated with RCTs also limit their usefulness in evaluating long-term impacts because it is costly to maintain the validity of the experiment over time.

Experimental evaluations are informative and offer a number of benefits; however, these issues, some of which are specific to curricular evaluation, suggest that a careful and rigorous non-experimental analysis can make a useful contribution to the literature. This is precisely what we provide in our study, using non-experimental data from the entire state of Indiana to estimate math curriculum effects on student achievement. We evaluate the three most-used curricula in the state from 1998 - 2004, which together, accounted for 86 percent of all curriculum adoptions in the grades that we study. Indiana provides the most detailed information about curriculum adoptions over time of any of the 50 states, and also provides thorough school- and district-level

[7] In the Agodini et al. (2009) study, the study team "provided logistical and financial support for any level of training the publishers indicated was appropriate." Although publishers typically provide training and support services whenever a new curriculum is adopted, they have added incentive to provide high-quality training and support during a RCT.

[8] In what is a relatively large-scale RCT, Agodini et al. (2009) evaluate four different curricula (more than the usual one or two curricula in other studies), but still only evaluate four school districts and 39 schools in the first wave of their study (in the second wave they will scale up to just over 100 schools). Their study was funded by the Institue for Education Sciences for over 21 million dollars. More typical RCTs are even more narrowly focused. Borman et al. (2008) and Resendez and Azin (2007) each evaluate just a single curriculum, looking across only five and four schools, respectively.

data about student achievement, demographics and school finances dating back many years. With the exception of the information about curriculum adoptions, similar data are available in many other states, suggesting that it would be straightforward to replicate our analysis elsewhere.[9]

We use school-level matching estimators in our study, adopting the pairwise-comparison approach suggested by Lechner (2002) to evaluate the three curricula. Drawing on the extensive methodological literature on matching, we show that the data conditions in Indiana are generally favorable to such an approach, particularly in our comparison of the two most popular curricula in the state. A key feature of our study is that we have constructed an extended data panel of Indiana schools containing information from multiple cohorts of students who were never exposed to the curricula that we evaluate. We use data from these cohorts to perform a series of falsification tests for our estimates, which show that our primary findings are unlikely to be driven by selection into the different curricula.

We highlight three primary results from our study: (1) differences across some math curricula have large short-term effects on student achievement, (2) as has been found with other educational inputs (see, e.g., Jacob et al., 2008; U.S. Department of Health and Human Services, 2010), math-curriculum effects do not persist over time, and (3) curriculum publishers that are relatively less effective in one adoption cycle do not lose market share in future adoption cycles. This latter result shares a common theme with prior research suggesting that educational administrators do not make optimal choices (Ballou, 1996). In this case, one explanation is the limited availability of reliable evidence on curricular effectiveness.

II. The Curriculum Selection Process

Curricula are adopted in Indiana for a single subject in each year across the entire state, and rotate in six-year cycles by subject. For example, Indiana's districts adopted new math curricula in 1998 and 2004, with an upcoming adoption in 2010. Similarly, recent reading

[9] It would not be expensive for states to track curriculum adoptions, particularly when compared to the costs of tracking some of the other information that is commonly collected.

adoptions occurred in 1994, 2000 and 2006. We focus our attention primarily on the math curriculum adoption that occurred in 1998.

The curriculum selection process in Indiana has centralized and decentralized components. The process begins with the state of Indiana's Department of Education (DOE) approving a list of selected curricula for use in the state. Upon receiving this list from the DOE, districts have three choices. First, and most commonly, they can adopt one or more of the state-approved curricula. Second, districts may choose to apply for alternate curricula that are not on the state-approved list, but this option is almost never used (e.g., no more than one out of the roughly 300 districts chooses this option in any grade in our data). Third, districts can apply for "continued use" where they continue to use the curricula that were adopted in the prior adoption cycle in that subject. The "continued use" option is not the same as a district choosing to adopt a new edition from the same publisher. Quite literally, it continues to use the old textbooks from the prior adoption cycle. Overall, over 98 percent of the districts in Indiana adopted new math curricula from the approved list during the 1998 adoption cycle in each grade.

We treat the DOE's approval process as exogenous to the districts, and focus our analysis on identifying differential curriculum effects among the curricula that are included on the DOE's approved list. The centralized approval process adds a constraint to the environment whereby we cannot evaluate curricula that are not approved by the state. However, it is not clear that the DOE's constraint is binding for districts in any meaningful way. For example, although districts can apply to use curricula outside the state-approved list, this rarely happens in practice, suggesting that most districts are content to choose among the available options. Perhaps more telling, the majority of the curriculum market share belongs to just a handful of publishers. Specifically, 86 percent of all curriculum adoptions in the grades that we study involve just three of the ten state-approved curricula during the adoption cycle of interest.[10]

[10] Indiana is one of 22 states that have a state-level component to the adoption process. Tulley (1989) finds that in states where there is not a centralized component to the adoption process, the curriculum review processes and lengths of use are similar despite the lack of a formal process dictating textbook choice. In conjunction with the

III. Data

We constructed a 17-year data panel of schools and districts in Indiana to evaluate the effects of math curriculum adoptions in grades one, two and three on grade-3 test scores in math (grade-3 is the first time that students are tested in Indiana). Among the 50 states, Indiana is the only state where curriculum-adoption information is available at the district level for multiple statewide adoption cycles.[11] Upon request, Indiana provides detailed school- and district-level information on test scores (from the Indiana state test, the ISTEP), attendance rates and enrollment demographics (including language minorities and students on free and reduced price lunch). Indiana also collects district-level financial information. Details on the district- and school-level information used in our analysis are provided in Table 1.

We evaluate the three curricula that dominated the market during the adoption cycle of interest (1998-2004). These curricula were published by Saxon, Silver-Burdett Ginn and Scott-Foresman, and they accounted for roughly 48, 23 and 15 percent of observed curriculum adoptions in the state, respectively. We denote the Saxon curriculum as curriculum *A*, the Silver-Burdett Ginn curriculum as curriculum *B*, and the Scott-Foresman curriculum as curriculum *C*.

Because we first observe student outcomes in grade 3, our estimates of curriculum effects characterize the impacts of *sequences of treatments*. That is, grade-3 test scores are presumably a function of the curricula to which students are exposed in grades one, two and three. To allow for cleanly identified curriculum effects, we exclude districts that adopted different curricula in different grades from our analysis. To illustrate the assignment problem for these districts, consider a district that adopted curriculum *A* in grade one and curriculum *B* in grades two and three. In identifying the effect of curriculum *A* relative to curriculum *B*, the schools in this district are not well-defined as either treatments or controls.[12]

---

limited practical importance of the centralized constraint, this suggests that the centralized component to Indiana's curriculum adoptions should not affect the generalizability of the results.

[11] In fact, in many states, the DOE does not even have a centralized database indicating which curricula are being used by districts within the state during the *current* adoption cycle, let alone historical information.

[12] Although we want to distinguish our estimates from estimates of single-year curriculum effects, our analysis is not related to the literature on sequences of treatments that also involve sequential decisions (see, for example, Lechner,

We also exclude districts where more than one curriculum was adopted because the data do not indicate which curricula were used by which schools within each district. Only in cases where a district adopted a single curriculum at all schools can we be sure that our treatment and comparison schools are properly identified.

We refer to districts that used the same curriculum at all schools and in all three grades as "uniform curriculum adopters." Restricting our analysis to these districts reduces our district sample size by eight percent and the analogous school sample size by seven percent (complete details on how we arrived at our final data sample are provided in Appendix Table A.1). That is, most districts are "uniform adopters." Overall, our analysis includes data from 213 districts and 716 schools. Contrasted with the experimental literature, where studies often focus on just a handful of schools and districts, our non-experimental design allows for a much broader evaluation of curricular effectiveness.

In Table 1 we report differences in means across the schools and districts that adopted the different curricula, using pre-adoption information from 1997. There are only small differences in test score performance and attendance outcomes across adopters of the different curricula, suggesting that selection into the curricula may be limited. However, there are noticeable differences in terms of school demographics, district size, and to some extent, median household income (measured at the district level from the US Census). Among other things, Table 1 indicates that Saxon adopters are disproportionately rural districts, as evidenced by their much smaller district sizes (and corresponding revenues) and their larger shares of white students.

IV. Curriculum Descriptions

In 1998, Mathematically Correct (MC), a national organization of mathematicians, scientists and engineers, qualitatively evaluated eight grade-2 math curricula, including the three curricula that we evaluate here. The MC evaluations were sponsored by the Texas Public Policy Foundation, a non-profit, non-partisan research institute. We briefly highlight the key differences

2004; Lechner and Miquel, 2009). In our study, districts make a treatment decision at a single point in time. Thus, methodologically, our evaluation procedure is the same as in the typical one-shot treatment case.

between the Saxon, Silver-Burdett Ginn and Scott-Foresman curricula as indicated by MC. We also report the MC rating for each curriculum, which was based on a 5-point scale (all three curricula received a similar rating from MC).

Curriculum A: Saxon Math (overall rating: 3.6)

The MC evaluation indicates that the program design is "easily implemented by teachers," and instructions to teachers are "clear and direct." In fact, the teacher's manual even includes scripted statements and questions for the teacher to ask to the class. The worksheets that students use are not necessarily related to the daily lesson, and contain a mixture of topics from prior lessons. One side of the worksheet is completed in class and checked, and the other side is assigned for homework. Oral assessments are given to individual students every 10 lessons, and are conducted while other students work on written work. Written assessments occur after every five lessons.

Saxon Math is very thorough in the topics that are covered, but more advanced topics are generally not covered. That is, this program supports learning effectively to a certain level but beyond that, achievement will be "very limited." As one example, of the three curricula of interest here, Saxon math is the only curriculum that does not cover addition and subtraction with three-digit numbers in the second grade. Overall, the MC evaluation suggests that Saxon Math may be the most effective curriculum for low-achieving students given its thorough coverage of the topics it covers, but will be less effective for high-achieving students.

Curriculum B: Silver-Burdett Ginn Math (overall rating: 3.4)

The teacher's manual provides guidance to teachers, although the guidance is not as direct as in Saxon Math. The teacher is given some discretion over how to present the material. In the example from the MC review, the teacher has two presentation choices for the lesson that are described as "visual/spatial learning" and one presentation choice that is described as "kinesthetic learning." In some cases, there is also a technology-based alternative. Student

worksheets are tied to the daily lesson. No information is given about the regularity of assessments or homework assignments.

The MC review highlights that this curriculum relies heavily on graphics to aid in the calculations (i.e., uses a "models" approach). In teaching addition and subtraction, for example, the curriculum relies heavily on graphics at first, but phases out their use later on. MC identifies the reliance of this curriculum on graphics as a weakness; however, notably, MC still rated curriculum *B* similarly to the other curricula in our study.

The level of this curriculum appears to be higher than that of Saxon Math – MC reports that students using this program have a "reasonable chance of moderate achievement levels" but also that the program is "not seen as supporting high achievement levels."

Curriculum C: Scott-Foresman Addison Wesley Math (overall rating: 3.8)

The teacher's edition received mixed reviews from the MC evaluation. Like the Silver-Burdett Ginn curriculum, the lessons also involve some discretion for teachers in terms of the activities that they use to teach each lesson (although there appear to be fewer teacher choices). Vocabulary development is an important part of this curriculum – new vocabulary words are introduced at the beginning of each lesson, and a verbal skills assessment occurs after each lesson. A one page homework sheet is also attached to each lesson.

The level of this program appears to be somewhere in between the levels in the prior two curricula. On the one hand, the MC review indicates that "the level is low in a few topics" and "at the top level of students…some topics should be augmented." On the other hand, the review also notes that "some areas are very well taught and at an excellent level."

It is important to note that while the MC reviews provide useful insights, they are not based on analyses of actual implementation, let alone student outcomes. We present the descriptions simply to highlight some of the differences that exist in organization, content, and presentation across these three math curricula.

V. Methodology

We use school-level matching estimators to identify curriculum effects. Matching is an increasingly common technique employed in empirical work, and the conditions under which matching will identify causal estimates of treatment effects have been well-documented (see, for example, Rosenbaum and Rubin, 1983; Heckman et al., 1997). The key benefits of matching relative to simple regression analysis are (1) matching imposes weaker functional form restrictions and (2) matching resolves any "extrapolation" problems that may arise in regression analysis by limiting the influence of non-comparable treatment and control units in the data (Black and Smith, 2004).

Briefly, the key assumption under which matching will return causal estimates of treatment effects is the conditional independence assumption (CIA). The CIA requires that potential outcomes are independent of the curriculum uptake decision conditional on observable information. Denoting potential outcomes by $\{Y_0, Y_1,..,Y_K\}$, curriculum treatment options by $T \in \{0,1….K\}$, and X as a vector of observable school- and district-level information, the CIA in our multi-treatment context can be written as:[13]

$$Y_0, Y_1,..., Y_K \perp T \mid X \qquad (1)$$

Conditional independence will not be satisfied if there is unobserved information that influences both treatment and outcomes. For example, if districts have access to information that is unobserved to the econometrician, Z, such that $\mathrm{P}(T = k \mid X, Z) \neq \mathrm{P}(T = k \mid X)$, and the additional information in Z also influences outcomes, matching will produce biased estimates of curriculum effects.

We estimate average treatment effects (*ATE*s) for the three curricula using a basic pairwise-comparison approach, as suggested by Lechner (2002), and match schools using an estimated propensity score (Rosenbaum and Rubin, 1983). In a comparison between curriculum *j*

[13] The CIA is actually a stronger assumption than is required to identify causal treatment effects, although it is difficult to imagine an environment where only the weaker but necessary condition of conditional *mean* independence is satisfied (Heckman et al., 1997; Imbens, 2004).

and curriculum m, and defining P_j as the probability of choosing j, we match schools by $(\frac{P_j}{P_j + P_m})$, where P_j and P_m are estimated using a multinomial probit that simultaneously models the three treatment options (Lechner, 2002).

We use kernel and local-linear-regression (LLR) matching estimators. These estimators construct the match for each "treated" school using a weighted average over multiple "control" schools, and vice versa. For a given curricular comparison, defining j as the treatment and m as the control, we estimate $ATE_{j,m}$ using the following formula:

$$\hat{\theta}_{j,m} = \frac{1}{N^S}[\sum_{j \in N_j \cap S_p} \{Y_j - \sum_{m \in I_{0j} \cap S_p} W(j,m) Y_m\} - \sum_{m \in N_m \cap S_p} \{Y_m - \sum_{j \in I_{0m} \cap S_p} W(m,j) Y_j\}] \qquad (2)$$

In (2), N^S is the number of schools using j or m on the common support, S_p. I_{0j} indicates the set of schools that chose m in the neighborhood of observation j, and I_{0m} indicates the set of schools that chose j in the neighborhood of observation m (where neighborhoods are defined based on propensity scores using a bandwidth parameter – see Appendix B). Y_j and Y_m are outcomes for treated and control schools, respectively, and $W(j,m)$ and $W(m,j)$ weight each comparison school outcome depending on its distance, in terms of estimated propensity scores, from the observation of interest. We omit a more detailed discussion of kernel and LLR matching estimators for brevity. For more information, see Heckman et al. (1997, 1998), and Fan (1993).[14]

Our matching estimators are conditioned on all of the observable information detailed in Table 1. *Ex ante*, it is unclear how unobserved selection into the different curricula might bias our estimates. For example, we might be worried that the adopters of the different curricula have student populations that differ in unobservable ways, or that differences in administrator quality

[14] Our results are robust to alternative matching estimators, and weighting estimators based on propensity scores (for discussions of weighting estimators see Imbens 2004; Millimet and Tchernis, 2009). See Section VIII for more detail.

that are correlated with curriculum adoptions may bias our results.[15] More generally, a limitation of matching analyses is that the CIA is not a testable assumption. However, we can get some indication as to the plausibility of conditional independence in our case by estimating curriculum "effects" for cohorts of students who were never actually exposed to the curricula of interest. If our matching procedure is producing estimates that are not biased by unobserved selection, we should estimate effects of *zero* for these cohorts. In Section IX, we present 80 different falsification estimates along these lines. These tests show that our primary findings are unlikely to be driven by selection on unobservables.

Finally, average treatment-on-the-treated effects (*ATT*s) may also be of interest. *ATT*'s can provide important information if the curricula differentially affect different subgroups of schools. For example, consider a case where $\theta_{j,m} = 0$. This could occur even if schools that chose *j* were better off for having chosen *j*, and schools that chose *m* were also better off for having chosen *m*. In addition to our *ATE* estimates, we also estimate *ATT*'s for all of the curricular comparisons and in both directions (that is, for each comparison we estimate both $ATT_{j,m}$ and $ATT_{m,j}$). We briefly discuss our findings in Section VIII, but in general, we gain little additional insight by estimating the *ATT*s.

VI. Timing and Treatment Definition

Timing is an important issue in our analysis. Our data panel spans 17 years, starting with the 1991-1992 school year and ending with the 2007-2008 school year. The curricula of interest were adopted in the fall of 1998, and replaced with new curricula in the fall of 2004. We observe seven cohorts of grade-3 students who were never exposed to the curricula of interest during the pre-period (1991-1992 through 1997-1998), one cohort that was exposed to the curricula in grade three only (1998-1999), one cohort that was exposed in grades two and three only (1999-2000), four cohorts that used the curricula in grades one, two and three and were thus "fully exposed" (2000-2001 through 2003-2004), one cohort that was exposed in grades one and two only (2004-

[15] Another concern would be that students may move across districts in response to curriculum adoptions. In results omitted for brevity, we find no evidence of such movement.

2005), one cohort that was exposed in grade one only (2005-2006), and two additional cohorts that were never exposed to the curricula in the post-period (2006-2007 and 2007-2008).

Recall that the estimated curriculum effects are based on grade-3 test scores, and as such represent the effects of sequences of treatments (T_{g1}, T_{g2}, T_{g3}). For the fully-exposed cohorts, the sequences for treatment and control schools are fully observed, meaning that these cohorts provide our cleanest estimates of curricular effectiveness. For the partially-exposed cohorts (the cohorts that were exposed to the curricula for at least one year, but less than three years), we can still estimate treatment effects because part of the curriculum sequence is observed. However, our analysis for these cohorts is limited because we do not observe curriculum treatments outside of the adoption cycle. A similar concern regarding out-of-cycle curriculum adoptions is relevant for our falsification tests (using the cohorts prior to 1998-1999, and after 2005-2006). This issue will be addressed in more detail in Sections VIII and IX when we present our results.

An additional concern related to timing is that the exposure levels of the different cohorts overlap with "curriculum familiarity" at schools. For example, the 1999-2000 cohort was exposed to the curricula for just one year, in grade 3, which was the year in which the curricula were first introduced at districts, and perhaps more importantly, to teachers. The 2005-2006 cohort was also exposed to the curricula for just one year, in grade 1, but at that point the curricula had been in use for many years. If teacher familiarity with curricula affects achievement, and curricula differentially affect outcomes depending on the grade level at which students are exposed, our estimates of curricular effectiveness at different grade-levels of exposure will be intertwined with the effects of different levels of curriculum familiarity. We can separately identify familiarity effects only across the four fully exposed cohorts - these cohorts were exposed for all three grades but differ with respect to their instructors' familiarity with the curricula.

Finally, a third timing issue involves district restructuring over the course of the 17 years of our data panel. Specifically, there is a pattern of school consolidations in the data such that the number of individual elementary schools decreases over time. As will be discussed in the

following section, we match schools based on their static characteristics from the 1996-1997 and 1997-1998 school years. School consolidations may alter the populations of students served by the schools that remain in operation over time. This will reduce the quality of our matches, and potentially introduce bias into our estimates.

In order for the school consolidations to bias our estimates they must be correlated with curriculum adoptions. However, this does not appear to be the case. Using a χ^2 test for independence, we fail to reject the null hypothesis that curriculum adoptions are independent of whether a district experiences a school closing (p-value $\approx$ 0.40). Additional evidence that our results are unlikely to be biased by school consolidations is provided in the next section where we evaluate the balance of the covariates across matched treatment and control schools over the entire course of the data panel. If the schools that drop out of our sample over time systematically adopted specific curricula, we should find that our treatment and control samples become less balanced as we move away from the matching years (1996-97 and 1997-98). We find little evidence of this, which further supports our contention that school closings are not correlated with curriculum adoptions (see Table 2).[16]

Although we do not expect the school consolidations to bias our results, they will reduce the quality of our matches as we move away from the 1996-1997 and 1997-1998 school years in the data panel. This will add noise to our estimates. Ultimately, we simply report this issue as a caveat, and caution the reader to interpret results that are estimated far away from the matching years more liberally. In an omitted analysis, we also considered a more direct solution to this problem – at any point where a school closing was observed in a district, we dropped all school-level observations from that district for the remainder of the data panel.[17] This alternative

[16] Of course, this use of balancing to test for non-random attrition will only catch non-random attrition if it is correlated with observables. Nonetheless, in conjunction with the our general test for independence between school closings and curriculum adoptions, the time-invariant balancing in our data further supports the contention that the school closings will not bias our estimates.

[17] We also performed an analogous procedure for schools that existed in 1996-1997, but came into existence between 1991-1992 and 1996-1997. If school closings re-shuffle student populations within districts, such an approach will reduce the number of bad matches in the data. There is enough natural variation in the enrollment data that we cannot always identify which specific schools are affected by a school closing, particularly when the closing

approach produces estimates that are qualitatively similar to what we report below, although the efficiency costs associated with discarding data from entire districts may be higher than those from allowing the less accurate matches to occur.[18]

VII. The Propensity Score

We use a multinomial probit (MNP) specification to estimate the pairwise propensity scores that we use to match schools. The covariates that we include in the MNP are documented in Table 1, and contain both school and district level information. At the school level, the propensity-score model includes controls for enrollment, demographics (race, free and reduced-price lunch status, language status) and outcomes (i.e., grade-3 test scores in math and language arts, and attendance) from the 1996-1997 school year, and controls for enrollment and demographics from the 1997-1998 school year (for brevity, means are not reported in the table for the 1998 information). At the district level, the model includes enrollment, outcome and finance controls from 1996-1997, and enrollment and finance controls from 1997-1998. We also use district-level zip codes to assign Census measures of local-area socioeconomic status to each school. Namely, we include controls for median household income and the share of the adult population who do not have a high-school diploma, both obtained from the year-2000 census. We treat these census variables as fixed area characteristics.

The covariates in the MNP specification were selected based on the process by which the curricula were adopted, with the objective of replicating the relevant information set available to schools and districts at the time of the adoption decision. The curriculum-adoption process in Indiana lasts approximately 18 months, and for the 1998 adoption this process culminated with a

---

school is small. Therefore, the most straightforward solution is to drop all schools in the district where the school closing is observed.

[18] An additional problem with this alternative strategy is that when a school closes we cannot be certain that the only other schools that are affected are in the same district. For example, if the closing school is on the border of another district, its students may change districts, in which case the district-dropping procedure would be doubly harmful – it would retain the schools in the new district where the students from the closed school were infused, and drop the schools from the district where the school closed even though these schools were not affected by the closing.

final decision in the summer of 1998.[19] Spring 1998 test scores would not have been available to decision makers prior to the adoption decision, and therefore, we do not include test scores from the 1997-1998 school-year in the MNP (we also omit annual attendance figures from 1997-1998 for the same reason). However, our findings are not qualitatively sensitive to reasonable adjustments to the MNP specification, including the addition of the 1997-1998 outcome variables. Similarly, our findings do not depend on whether we include additional years of lagged test scores in the propensity-score specification. An important reason for limiting the number of lagged years of achievement in the propensity-score specification is that we want to use as many years of data as possible for the falsification tests. Each year of data that we use to match schools is one less year that we can use in the falsification exercise.

In each comparison we match treatment and control schools based on the estimated pairwise propensity scores, and test for balance in the covariates among the treated and control samples used for estimation.[20] Balancing tests are motivated by Rosenbaum and Rubin (1983). The tests determine whether $X \perp T \mid P(T = k \mid X)$, a necessary condition if the propensity score is to be used to reduce the dimensionality of the matching problem to one.

Although achieving covariate balance is important for any matching analysis that relies on a propensity score, there is no clearly preferred test for balance. Furthermore, in some cases, different balancing tests return different results (Smith and Todd, 2005). Given this limitation, we consider two different tests. The first is a regression-based test suggested by Smith and Todd (2005), estimated separately for each covariate in each year of our analysis:

$$\begin{aligned} X_k = {} & \beta_0 + \beta_1 \hat{P}(X) + \beta_2 \hat{P}(X)^2 + \beta_3 \hat{P}(X)^3 + \beta_4 \hat{P}(X)^4 \\ & + \beta_5 T + \beta_6 * T * \hat{P}(X) + \beta_7 * T * \hat{P}(X)^2 + \beta_8 * T * \hat{P}(X)^3 + \beta_9 * T * \hat{P}(X)^4 + \varepsilon \end{aligned} \tag{3}$$

[19] The timeline for the current math-curriculum adoption cycle is available at http://www.doe.in.gov/olr/docs/CHRONOLOGYFORTHE2009MATHEMATICSADOPTIONApr09.pdf.

[20] For brevity we do not report the results from the propensity-score model, but they are available upon request. To provide a sense of the predictive power of the covariates in the model, we estimate separate linear-regression models for each curriculum comparison where the dependent variable indicates the adoption of one of the curricula, and the independent variables are the covariates from the MNP. Within comparison pairs, the covariates explain 23 to 42 percent of the variability in curriculum adoptions.

In (3), X_k represents a covariate from the propensity-score specification, $\hat{P}(X)$ is the estimated pairwise propensity score, and T indicates treatment. We test whether the coefficients $\beta_5 - \beta_9$ are jointly equal to zero in each regression – that is, we test whether treatment predicts the X's conditional on a quartic of the propensity score.

The second test measures the absolute standardized bias in observables after matching, and was originally suggested by Rosenbaum and Rubin (1985). The formula for the absolute standardized bias for covariate X_k is given by:

$$SDIFF(X_k) = \frac{\left| \frac{1}{N^S} \left[\sum_{j \in N_j \cap S_p} \{X_{kj} - \sum_{m \in I_{0j} \cap S_p} W(j,m) X_{km}\} - \sum_{m \in N_m \cap S_p} \{X_{km} - \sum_{j \in I_{0m} \cap S_p} W(m,j) X_{kj}\} \right] \right|}{\sqrt{\frac{Var(X_{kj}) + Var(X_{km})}{2}}} * 100 \qquad (4)$$

The numerator in (4) is analogous to the formula for our matching estimators in (2) where we replace Y with X_k and take the absolute value. Note that the variances in the denominator are calculated using the full sample (i.e., they include observations that are not on the common support). A weakness of the standardized-bias approach is that there is not a clear rule by which to judge the results, although Rosenbaum and Rubin (1985) suggest that a value of 20 is large.

Our MNP specification uses 32 school- and district-level covariates. The results from the balancing tests are reported in Table 2 by comparison and year. From the regression tests we report the number of covariates where the F-tests reject the null hypothesis at the 5- and 10-percent levels (the former group is a subset of the latter), and the average p-values across all F-tests. We also report the average absolute standardized bias across all covariates. The median absolute standardized bias is also often reported – we omit it for brevity but note it is always smaller than the average.

Table 2 indicates that our comparison between *B* and *A* achieves better balance than our other comparisons. For this comparison, both the regression tests and the standardized-bias results suggest that schools are well-matched. For our comparisons between *C* and *A*, and *C* and *B*, the covariates appear to be less balanced, although it is not clear that the levels of imbalance

in these comparisons are cause for concern. For example, the average p-values from the F-tests in both comparisons are fairly close to 0.50 in all years, which suggests good balance, despite there being more unbalanced covariates than would be expected by chance in both cases. Similarly, although the average absolute standardized bias is much larger in these comparisons than in our comparison between *B* and *A,* by some standards it is still quite reasonable.

Although it is not obvious that the level of imbalance in any of our comparisons is large enough to be problematic, in unreported results we considered many alternative propensity score specifications where we added various combinations of higher-order and interaction terms to the MNP in an effort to improve covariate balance. Likely due in part to our relatively small treatment and control samples (compared to the general matching literature), these alternative models generated only modest improvements in covariate balance, and did not affect our findings qualitatively. Therefore, we proceed below using the simple MNP to predict curriculum adoptions for each comparison, noting that the balancing results are less compelling in our comparisons between *C* and *A*, and *C* and *B*.

We also calculate the divergence between the densities of the estimated propensity scores for treated and control units in each comparison. Intuitively, density divergence will affect the precision of the estimates obtained from matching. Density divergence has been discussed in numerous studies, including Frölich (2004), who measures divergence using the Kullback-Leibler (KL) information criterion. We follow his approach here, using kernel-density plots based on the Epanechnikov kernel. For example, denoting $\rho_{21} = (\frac{P_B}{P_B + P_A})$ as the probability of choosing *B* over *A*, we estimate the divergence between the densities of ρ_{21} for treated and control units as:

$$KL = \int \ln(\frac{f_{p|T=B}(\rho_{21})}{f_{p|T=A}(\rho_{21})}) f_{p|T=B}(\rho_{21}) d\rho_{21} + \int \ln(\frac{f_{p|T=A}(\rho_{21})}{f_{p|T=B}(\rho_{21})}) f_{p|T=A}(\rho_{21}) d\rho_{21} \quad (5)$$

In (5), $f_{p|T=B}(\rho_{21})$ is the density function of ρ_{21} among schools treated with *B*, and $f_{p|T=A}(\rho_{21})$ is the analogous density function for schools that used curriculum *A*. A KL-information-criterion measure of zero suggests that the densities are identical, and the measure increases with density divergence. Note that when the parameters of interest are average effects of treatment on the treated, researchers use a unidirectional version of the KL information criterion (see, for example, Frölich, 2004). In our case, where average treatment effects are the parameters of interest, we use the bidirectional information criterion originally suggested by Kullback and Leibler (1951).

Figure 1 plots the estimated density functions of the propensity scores for treatment and control schools for each pairwise comparison, and Table 3 reports the corresponding KL information criteria. Similarly to the balancing tests, the density-divergence measures suggest that the data conditions are most favorable in our comparison between *B* and *A*. Density divergence is largest in our comparison between *C* and *A*.[21]

Both the balancing tests and the density-divergence measures indicate that our data are best-suited to compare curricula *B* and *A*, which combined, accounted for over 70 percent of the curriculum market in Indiana during the 1998 adoption cycle. In the other two comparisons the data conditions are generally less favorable; however, even in these comparisons, it is not clear that they are cause for concern. We consider the reliability of our results from each comparison in more detail when we present the falsification tests in Section IX.

VIII. Estimates of Curricular Effectiveness in Math

Rather than overwhelm the reader with estimates using the numerous matching algorithms available in the literature, we instead present estimates using kernel and local-linear-regression (LLR) matching only (for details on these and other matching estimators, see, for

[21] Frölich (2004) uses unidirectional density divergence measures to describe the data conditions in his evaluation of the performance of different matching estimators. Although the one-sided measures are not directly comparable to the two sided measures; roughly speaking, our comparison between *B* and *A* corresponds to Frolich's most-favorable density design, our comparison between *C* and *B* his middle design, and our comparison between *C* and *A* his least favorable design. This is purely by coincidence.

example, Heckman et al., 1997, 1998; Mueser et al., 2007). Frölich's (2004) analysis indicates that kernel matching in particular should perform well in our context.[22] As for LLR matching, the evidence in the literature is mixed.[23] Although our estimates using LLR matching are less precise than the kernel-matching estimates, they are generally very similar. We present results using the Epanechnikov kernel for both types of matching estimators. In unreported results available upon request we show that our results are robust to alternative estimators, including kernel and LLR matching estimators that use the Gaussian kernel, other matching estimators based on simple pair matching or radius matching using various radii, and regression-adjusted and weighting estimators.

Table 4 presents results for all grade-3 cohorts who were ever exposed to the curricula of interest using fixed-bandwidth matching estimators where the bandwidths are obtained via conventional cross-validation.[24] All of our matching estimators impose the common support condition. We also report OLS estimates where we regress test score outcomes on the covariates used in the propensity score model and indicator variables for curriculum adoptions, retaining our pairwise comparisons (that is, when we compare *B* to *A*, we drop all schools at districts that adopted *C*). The standard errors for the matching and OLS estimates are clustered at the district level and the matching-estimator standard errors are bootstrapped using 250 repetitions.[25] We obtain the optimal numbers of bootstrap repetitions to use for our standard error calculations following Ham et al. (2006), who use a special case of Andrews and Buchinsky (2001).[26]

---

[22] Frölich's (2004) study also suggests that ridge matching should perform well, but the ridge parameter will lead to bias in the case of multiple covariates (Frölich uses a single-covariate setting). See Heckman et al. (1998) for details.

[23] For instance, Fan (1993) indicates that local linear regression has better sampling properties than the standard kernel estimator, and Caliendo and Koepinig (2005) suggest LLR is particularly useful when controls are distributed asymmetrically around treated observations. Frölich (2004) notes that LLR matching will perform worse in regions of sparse data, which is consistent with the large standard errors that we estimate in some years using LLR matching in our comparisons with less density overlap.

[24] In some cases the cross-validation estimates of the loss function are fairly flat. In these cases, we combine "visual inspection" with cross-validation to choose the optimal bandwidth. See Appendix B for details.

[25] Abadie and Imbens (2006) show that bootstrapping methods cannot be used to obtain standard errors for nearest-neighbor matching estimators, but their study does not apply to smoother estimators like those used here.

[26] For our estimators, the optimal number of bootstrap repetitions is consistently near 200. We use 250 repetitions to ensure that we meet or exceed the optimal repetition count in each year.

Each cohort is labeled in the tables according to the year of its spring test score (i.e., the 1998-1999 cohort is labeled "1999"). Recall that the 1999, 2000, 2005, and 2006 cohorts were only partially exposed to the curricula, while the cohorts from 2001 through 2004 were exposed for all three years. All of the effects in the table are standardized using the distribution of school-level test scores. For example, the estimate in Table 4 for $ATE_{B,A}$ in 2002 indicates that, among the sample of schools that chose *B* or *A,* the average effect of using *B* instead of *A* was 0.40 standard deviations of the distribution of school-level math test scores. More typically, researchers report effects that are standardized based on the distribution of *individual-level* scores, but we do not have access to the distributions of individual-level scores over the entire course of the data panel (specifically, we do not have these distributions for the years prior to 1999-2000). In Appendix Table A.2, for each year where we have access to the individual-level distribution of test scores (such that we could compute the standard deviation), we provide the scaling factors that convert the estimates in Table 4 into the more common metric. Roughly speaking, dividing the coefficients by three returns estimates in the metric of standard deviations of the individual-level distribution of scores.

Focusing first on our largest comparison between *B* and *A*, and the estimates for the fully-exposed cohorts (2001 – 2004), we find that curriculum *B* meaningfully outperformed curriculum *A*. Averaging the kernel-matching estimates across all four fully-exposed cohorts, and using the appropriate scaling factors in Appendix Table A.2, the effect of using curriculum *B* instead of *A* was approximately 0.12 standard deviations of the test. Our estimates are also consistent with *C* outperforming *A*. There we estimate an average effect of roughly 0.06 standard deviations of the student-level distribution of scores, although only two of the four estimates are statistically significant and the estimate from 2004 is particularly small. Our results also suggest, at least weakly, that *B* outperformed *C*, although inference from this comparison is limited because the estimates are imprecise.[27]

[27] We also note that our inability to follow individual students over time implies some downward bias in our estimates to the extent that students switch curricula between grades one and three. For example, even among

The magnitudes of the estimated curriculum effects are economically meaningful, particularly when weighed against the marginal costs associated with choosing one curriculum over another. Fryer and Levitt (2006) show that between grades one and three, the black-white achievement gap grows at a rate of approximately 0.10 standard deviations per year.[28] In our most-compelling curriculum comparison, we estimate that the effect of choosing curriculum *B* over curriculum *A* is roughly equivalent to one year's worth of expansion of the black-white achievement gap. Given that the curricula are very similarly priced (the texts from *A, B* and *C* were, averaged over grades 1-3, $23.08, $24.80 and $25.34 each, respectively, in 1998 dollars), selecting a better curriculum appears to be a cost-effective way to improve student achievement.

We do not find any evidence of curriculum-familiarity effects for the fully-exposed cohorts. If curriculum familiarity were important for teachers, we might expect the 2001 and 2002 cohorts, for example, to be less affected by curriculum differences than the cohorts in 2003 and 2004 (under the assumption that when familiarity is low, curriculum implementation by teachers reverts toward a common mean). There is no evidence of such a trend in curricular effectiveness across cohorts.

Our results for the partially-exposed cohorts differ by comparison. One common theme is that the point estimates for the 2005 and 2006 cohorts are generally larger than for the 1999 and 2000 cohorts. In fact, in our comparison between *B* and *A*, the estimates for the 2005 and 2006 cohorts are large and statistically distinguishable from zero. Note that the 2005 and 2006 cohorts were exposed to the curricula as the treatments were winding down, while the earlier cohorts were exposed when the treatments were just beginning. If using the respective curricula for multiple years affects schools and teachers, regardless of student exposure, this may explain our findings for the 2005 and 2006 partially-exposed cohorts. Also of interest is that, per Section IV, curriculum *B* is distinguished from the other curricula in terms of mathematics pedagogy

---

students from the fully-exposed cohorts, across-district movers who are tested in grade-3 may only be partially exposed to their assigned curricula.

[28] Fryer and Levitt (2006) analyze a different testing instrument; however, similar estimates of the black-white achievement gap spread are available elsewhere (see, for example, Chubb and Loveless, 2002).

(specifically, curriculum *B* relies more on mathematical models). Although it would be entirely speculative to link the benefits of curriculum *B* to any specific attribute, this pedagogical difference would have the potential to stay with teachers and administrators beyond the curriculum adoption cycle of interest.[29]

In terms of gaining further inference from the partially exposed cohorts, we face two obstacles. First, although we cannot distinguish any curriculum-familiarity using the fully-exposed cohorts, there may be familiarity issues upon immediate adoption, which would affect the 1999 and 2000 cohorts but not the 2005 and 2006 cohorts. Second, the students in all of the partial-exposure cohorts were exposed to other curricula in other adoption cycles. This is likely to attenuate the partial-exposure estimates. The degree of attenuation will depend on the extent to which curricular quality is correlated across adoption cycles for treatment and control schools, which we explore to the extent possible in Table 5.

Table 5 compares curriculum adoptions in the 2004 adoption cycle across uniform adopters from 1998 (recall that we do not have curriculum adoption data from the prior cycle in 1992). For brevity, the table shows adoption shares in 2004 only for the four most popular curricula from that adoption cycle (published by Saxon, Harcourt, Houghton-Mifflin and Scott-Foresman). For the 2005 and 2006 cohorts, Table 5 provides direct information about the curricula to which they were exposed after the 1998 adoption cycle. For the 1999 and 2000 cohorts, it is merely suggestive about the extent to which curriculum adoptions in the prior cycle may have been correlated with the 1998 adoptions. The table shows that while Saxon adopters (curriculum *A*) in 1998 were much more likely to adopt Saxon in 2004, adopters of the other two curricula are quite dispersed across alternative options during the 2004 adoption cycle. Without knowing the respective qualities of the different curricula adopted outside of the 1998 adoption cycle, including those from the same publishers (there is no evidence that we are aware of on the

[29] Equally interesting is that Mathematically Correct, which has a preference for non-model-based mathematics instruction, rated curriculum *B* similarly to *A* and *C*. Given Mathematically Correct's pedagogical preference, it is almost certain that curriculum *B* was downgraded for using models, which suggests the quality of the curriculum outside of this issue may be high.

persistence of publisher quality), it is difficult to form expectations based on the patterns in Table 5.[30] Ultimately, given the potential for attenuation in the estimates for the partially exposed cohorts, and the sizes of our standard errors, we cannot make strong inference about partial-exposure curriculum effects.

Table 5 is also informative about the changing market shares of curriculum publishers over time. It shows that the publisher of curriculum *A*, despite its relative underperformance, maintained its near-50-percent market share from the 1998 adoption cycle to the 2004 adoption cycle. Although curriculum *B* was the most effective curriculum during the 1998 adoption cycle, it did not appear in 2004. The publisher of curriculum *B* was bought by Pearson Publishing during the 1998 cycle and Pearson phased out curriculum *B* in favor of curriculum *C*, which it also publishes. Curriculum *C's* market share fell from roughly fifteen to nine percent across adoption cycles.

Finally, in an omitted analysis we also considered the possibility that the treatment effects depend on treatment status. For example, despite our finding that curriculum *B* outperformed curriculum *A* on average, it could be that curriculum *A* was still better for schools that actually chose *A*, while curriculum *B* performed better for schools that chose *B*. To investigate the extent to which the curriculum effects might depend on treatment status, for each of our comparisons we estimated average treatment-on-the-treated effects (*ATT*) in each direction. Our findings provide few insights. In our comparison between *B* and *A*, the treatment effects do not depend on treatment status. Similarly, the *ATT*'s in our comparison between *C* and *B* do not suggest differential effects (although again, these estimates are noisy). Only in our comparison between *C* and *A* do we find any evidence of differential curriculum effects. There, curriculum *A* appears to perform less poorly relative to *C* at schools that actually chose *A*. Nonetheless, even our

[30] Evidence on the persistence of publisher quality would be difficult to obtain without the availability of consistent comparisons over time. For example, because Silver-Burdett Ginn did not offer a curriculum in Indiana during the 2004 adoption, our most reliable comparisons (per Section VII) cannot be replicated in the later adoption cycle. Even more, we cannot reliably compare Saxon and Scott-Foresman in 2004 because of the large decline in Scott-Foresman's market share across adoption cycles. Even in cases where curriculum publishers are consistently represented across adoption cycles, long cycle durations imply that long data panels will be required to evaluate the persistence of publisher quality.

estimates of $ATT_{A,C}$ suggest that schools that actually chose *A* would have been better off had they instead chosen *C*.

Overall, our most reliable estimates of curricular effectiveness come from the four cohorts of fully-exposed students who used the curricula of interest in grades one, two and three. Our estimates based on these cohorts indicate that curriculum *B* outperformed curriculum *A* by a substantial margin. We also find that *C* outperformed *A*, although the differential effect was smaller. The statistical imprecision associated with our comparison between *C* and *B* limits inference, but if anything, our estimates suggest that *B* outperformed *C*. The relative underperformance of curriculum *A* did not adversely impact the publisher's market share in the next adoption cycle in Indiana.

IX. Falsification Tests

Matching estimators will not return causal estimates if conditional independence is violated. Although conditional independence is not a testable assumption, we provide some evidence on its plausibility using a series of falsification tests. We present falsification tests based on data from students who were never actually exposed to the curricula of interest (e.g., cohorts of grade-3 and grade-6 students from the mid 1990s), and from students who were exposed, but we estimate curriculum effects on reading test scores. For the students who were never exposed to the curricula we expect to estimate "effects" that are statistically indistinguishable from zero. For the reading estimates, timing does not rule out the possibility of causal effects for some cohorts; however, at most, we would expect only small across-subject spillover effects.

Potentially confounding both types of falsification estimates are correlations in curriculum adoptions across grades, subjects, and adoption cycles. Recall from Table 5 that there are non-zero correlations in math-curriculum adoptions across adoption cycles. Not surprisingly, in unreported results (omitted for brevity and available upon request) we also find that math curriculum adoptions are correlated across grades within adoption cycles, and to a lesser extent, with curriculum adoptions in other subjects (where the adoptions overlap imperfectly with the

math adoptions – see Section II). The correlations between the curricula of interest and the other curricula to which the falsification cohorts were exposed could potentially confound the falsification tests. For example, if curriculum quality is correlated across adoption cycles for districts, the falsification estimates will capture more than just bias, making them difficult to interpret. However, in practice, the correlations in curricular quality across adoption cycles do not appear to be strong enough to limit inference from our falsification exercise - almost all of the falsification estimates are statistically indistinguishable from zero.

For brevity, we only report falsification estimates using kernel matching with the Epanechnikov kernel.[31] We present 80 falsification estimates in all (but note that the tests are not independent).[32] Summarizing the results, the tests do not uncover any consistent evidence of selection bias in any of our comparisons, although similarly to Table 4, the falsification estimates are noisy in our comparison between *C* and *B*, limiting inference.

We begin by estimating curriculum "effects" on math test scores for cohorts of grade-3 students from 1992 through 1996, and 2007 and 2008 (recall that we use data from 1997 and 1998 to match schools). The results are reported in Table 6. Our most-convincing falsification estimates are from the 1992-1996 cohorts, who passed through Indiana schools prior to the curriculum-adoption cycle we study. For these cohorts, all of the estimates are small and statistically indistinguishable from zero with the exception of the 1992 estimate in our comparison between *C* and *A*. Although the 2007 and 2008 cohorts were not actually exposed to the curricula of interest, their outcomes were observed after the curriculum-adoption cycle we study. This leaves open the possibility of non-zero treatment effects for these cohorts, limiting inference to some degree, but even so, none of the estimates from these cohorts are statistically significant.

[31] In unreported results we verify that our findings are robust to using the Gaussian kernel instead of the Epanechnikov kernel and to alternative matching estimators.

[32] If the falsification tests were independent we would expect roughly eight "false positives" in total. However, treatment and control schools are uniformly defined over time, making it unclear how many false positives to expect.

Next we estimate curriculum "effects" using cohorts of grade-6 students who were never exposed to the curricula of interest (cohorts from 1993-2001). For these falsification tests we use the same matching procedure to predict the same treatments (the uniform adoption of curriculum *A, B or C* in grades one, two and three), only we match schools that have grade-6 classrooms and estimate the "effects" of the curricula on grade-6 achievement.

Many districts teach grades three and six in different buildings (i.e., elementary and middle schools). Further, multiple elementary schools generally feed into a single middle school, meaning that the grade-6 samples of schools are much smaller than the grade-3 samples. This turns out to be problematic for our grade-6 comparisons involving curriculum C because our sample of curriculum-C districts is small. When we focus our attention on grade-6 schools, our sample of schools in curriculum-C districts falls to below 100 (roughly 80, on average, across the data panel), and we cannot balance treatment and control schools in either of our comparisons involving this curriculum. For example, taking the average p-values from the Smith and Todd (2005) balancing regressions across years for the comparisons involving curriculum *C*, they fall from roughly 0.50 in the grade-3 analysis (as reported in Table 2), to roughly 0.20 in the grade-6 analysis. Alternatively, in our grade-6 comparison between *B* and *A* the sample sizes are much larger, and the balance in the grade-6 comparison is only slightly worse than in the grade-3 comparison (the average p-value across years falls to just below 0.50 in the grade-6 sample).[33]

The lack of balance in the grade-6 comparisons involving curriculum *C* suggests that estimates from these comparisons will not be informative. Therefore, we present grade-6 falsification estimates only for our comparison between *B* and *A*. These estimates are reported in

[33] That the grade-6 comparisons involving curriculum *C* are unbalanced while the grade-3 comparisons appear to be roughly balanced is interesting but perhaps not surprising given the small samples of schools that are available for the grade-6 analysis. We can only speculate as to why the grade-6 schools are less balanced in the data. Our small samples are surely partly responsible, but how students matriculate through schooling levels at different districts, and how students are dispersed across sets of elementary schools that feed into a single middle school, may also be important. One easily observable difference between districts that adopted curriculum *C* and the other districts is that the ratio of grade-6 to grade-3 schools is smallest for curriculum-*C* districts. This is likely because curriculum-*C* districts tend to be larger (see Table 1). Ultimately, the grade-6 data in our comparisons involving curriculum *C* are not well-suited for a matching analysis.

Table 7, where we estimate one non-zero "effect" in 1993, but otherwise, the point estimates are generally small and statistically indistinguishable from zero.

In Table 8 we return to our well-balanced grade-3 samples and estimate math curriculum effects on reading test scores for all cohorts in the data panel. Students in the cohorts from 1992 through 1996, and 2007 and 2008, were never exposed to the curricula of interest. The other cohorts of students were exposed, and it is unclear *a priori* whether we should expect any across-subject spillover effects. We suggest four possible mechanisms that may generate spillover effects. First, math curricula may directly affect reading performance. As an example, math curricula may differentially use word problems, which could lead to differential effects on reading scores. Second, the training for teachers associated with each math curriculum could affect teacher performance in other subjects. Third, a better math curriculum may afford teachers more time to spend on reading instruction. Fourth, a better math curriculum may increase the return to math instruction and encourage teachers to substitute out of reading instruction and into math instruction. These latter two possibilities are analogous to income and substitution effects from basic microeconomic theory. The direction of the across-subject spillover will depend on which effect dominates.

Although we do not have a strong prior about whether math curricula affect reading outcomes, one straightforward expectation is that the effects of math curricula on math test scores should be larger in magnitude than their effects on reading test scores. Thus, at its most basic level, this final test should confirm this result. Table 8 presents estimates for the effects of the math curricula on reading test scores throughout our data panel, and indeed, the point estimates are generally small and there is only one statistically significant estimate (in our comparison between *B* and *A* in 2002).

While all of the estimates in Table 8 are relatively small, and only one is statistically indistinguishable from zero, taking the point estimates at face value at least raises the possibility that our primary findings are biased, particularly in our comparisons between *B* and *A*, and *C* and

B.[34] Therefore, we briefly consider how a pure-bias interpretation of the reading estimates would impact our results by assuming that across-subject spillover effects are zero. To do this, we estimate math-curriculum effects on schools' de-trended math test scores, where we de-trend each school's math score by separately standardizing its math and reading test scores, and subtracting the reading score from the math score. We omit the estimates for brevity, but note that they are in line with what would be expected by subtracting the stand-alone reading estimates from the stand-alone math estimates. Specifically, the estimates still indicate that *B* outperforms *A*, although the point estimates fall by roughly half, and that *C* outperforms *A*. The estimates that are statistically significant in Table 4 for these comparisons remain statistically significant in the de-trended analysis. In the comparison between *C* and *B*, the curricula are not statistically distinguishable in any year using the de-trended estimates. We treat the de-trended results as lower bounds because they assume that across-subject spillover effects are zero.[35]

X. Persistence

Finally, we use our extended data panel to evaluate the persistence of curriculum effects over time. Specifically, in our comparison between *B* and *A*, we ask whether the cohorts of students who were exposed to curriculum *B* in grades one, two and three still performed better by grade six.[36] We measure persistence using test score outcomes for cohorts of grade-6 students between 2002 and 2008. These cohorts correspond to the cohorts of grade-3 students who were exposed to the curricula of interest in our primary analysis – for example, the 2005 cohort of grade-6 students is also the 2002 cohort of grade-3 students. The fully exposed cohorts were in grade six between 2004 and 2007.

---

[34] One seemingly obvious source of bias is that we do not condition on reading curriculum adoptions in producing our estimates. However, math and reading curriculum adoptions are not highly correlated, and in an omitted analysis we show that this cannot account for our findings in Table 8.

[35] We compute effect sizes for our comparisons between *B* and *A*, and *C* and *A*, as in the previous section by averaging the de-trended estimates across the four fully-exposed cohorts. This suggests an effect size of choosing *B* over *A* of approximately 0.07 standard deviations of the test. The effect of choosing *C* over *A* remains the same, 0.06 standard deviations, by virtue of the small reading estimates.

[36] As discussed in the previous section, we are unable to construct observationally equivalent comparisons of treated and control schools from the grade-6 sample in our evaluations involving curriculum *C*. Therefore, we only examine persistence in our comparison between *B* and *A*.

Two issues merit attention in our persistence analysis. First, if there are test-score ceilings in higher grades on the Indiana test, it will be difficult to detect persistence effects because the tests in later grades may not adequately differentiate student learning. We test for ceiling effects following Koedel and Betts (2010) and find that the testing instruments should be sufficient to detect any persistence effects should such effects exist. A second concern is that we cannot track individual students over time in the data, and as a consequence our assignments to curriculum treatments during grades 1-3 may not be accurate for all students in any given cohort. That is, while every school that contains a grade-6 classroom is attached to a district, allowing us to identify the curriculum to which students would have been exposed in grades 1-3 if they attended a school in that same district, some students may have moved districts between grades 1-6. This churning implies that some of the students who contribute to a school's grade-6 test score were not actually treated with the district's curriculum in the early grades. This will add noise to our treatment classifications, attenuating any estimated persistence effects.[37]

Table 9 presents our persistence findings, again using kernel matching with the Epanechnikov kernel. Even though we expect our results to be attenuated to some extent per the previous discussion, the estimates in the table provide little indication that curriculum effects persist over time. Put differently, for the estimates in the table to be driven by downward bias from student movement across districts, the amount of student movement would need to be inordinately large. This result is consistent with a large body of evidence pointing to a general lack of persistence in the effects of educational inputs (see, for example, Jacob et al., 2008; U.S. Department of Health and Human Services, 2010), and raises doubts about the extent to which administrators can improve student performance in the long run by choosing more effective curricula.

[37] As noted above, student churning across districts is also a problem in our primary analysis, although less so. For example, if a student changes districts in grade-2, she may change curricula. Therefore, all of our estimates will be biased toward zero to some extent.

XI. Conclusion

We identify causal curriculum effects using non-experimental data from the state of Indiana. A key component of our study is our falsification exercise, where we use data from multiple cohorts of students who were never exposed to the curricula of interest, and students' out-of-subject test scores, to show that our findings are unlikely to be driven by selection into the different curricula. In cases where data conditions are favorable, and some form of confirmation exercise is possible (like our falsification tests), much can be learned from careful, non-experimental work. A caveat relating to curricular evaluation is that, somewhat surprisingly, most states do not centrally track curriculum adoptions. Given that it would be relatively inexpensive to track this information, and that curricula play such a large role in students' everyday learning experiences, this seems peculiar.

Currently, the bulk of the curricular-effectiveness debate is not based on rigorous evidence from analyses of implementation. For example, in addition to the general lack of rigor in comparative curricular evaluations (WWC, 2007), much of the literature relies on case studies, or content studies, where curriculum impacts on student outcomes are not measured (National Research Board, 2004). Rigorous scientific evidence about how different curricula actually affect student achievement is needed in order for administrators and educators to make informed decisions. Our study provides such evidence on a scale not yet seen in the curriculum-evaluation literature.

That our study is non-experimental allows us to bypass some of the limitations inherent to experimental analyses of curricular effectiveness. These limitations include the experimentation on a subpopulation problem (Manski, 1996), and the possibility of publisher Hawthorne effects. The latter concern seems particularly important given publishers' active roles in curriculum experiments. Another benefit of our non-experimental approach is that it is feasible to replicate in other environments both methodologically *and fiscally*. In contrast to the ongoing project by Agodini et al. (2009), a particularly well-designed RCT that is funded by the Institute

for Education Sciences for roughly 21 million dollars over five years, our study was performed using publicly available data at only a small fraction of this cost.

We also note several limitations of our study. For one, we do not have enough data, or the right kind of data (i.e., student level), to evaluate the extent to which curricula differentially affect different types of students (e.g., high and low-achieving, English-proficient and ESL, etc.). This deficiency in our analysis is likely to be less problematic in the future because districts and states continue to develop longitudinal databases that track individual students. These data could be linked to curriculum data, if such data were available, quite easily. We also depend on the standardized test administered by the state of Indiana as our outcome measure (the ISTEP). While we expect our results to extrapolate well to other states or districts that use similar tests, they may not carry over to states or districts where the testing instrument differs greatly in content. Our results also may not extrapolate well to states or regions where the student population differs greatly from the student population in Indiana, which is a fairly rural state.

Our findings indicate that students in Indiana who used curricula *B* or *C* outperformed students who used curriculum *A*. In our most compelling comparison, between *B* and *A*, the effect of exposure to the better curriculum for three consecutive years is roughly 0.12 standard deviations of the grade-3 ISTEP test. This effect is similar in magnitude to one year's growth of the black-white achievement gap over these grades (Fryer and Levitt, 2006). Interestingly, despite the consistent underperformance of curriculum *A* in our analysis, the publisher of curriculum *A* did not lose market share in the next curriculum adoption cycle in Indiana. There are many possible explanations for this finding, ranging from a lack of reliable information available to administrators about curricular quality (WWC, 2007), to poor decision making by educational administrators (also see Ballou, 1996).

Overall, our results are encouraging because choosing a better curriculum can non-negligibly improve student performance. Further, the near-zero marginal cost of choosing one curriculum over another suggests that implementing a better curriculum will be quite cost-effective. However, our finding that curriculum effects do not persist over time, although not

unique to curriculum in education, dampens enthusiasm about the potential benefits of improved curricula. By grade six, the benefits of the most-effective curriculum in our study are no longer distinguishable.

References

Abadie, Alberto and Guido W. Imbens. 2006. On the Failure of the Bootstrap for Matching Estimators. NBER Technical Working Paper No. 325.

Agodini, Robert and Barbara Harris and Sally Atkins-Burnett and Sheila Heaviside, and Timothy Novak. 2009. *Achievement Effects of Four Early Elementary School Math Curricula*. National Center for Education Evaluation and Regional Assistance, U.S. Department of Education, Institute of Education Sciences. NCEE 2009-4052.

Association of American Publishers. 2001. *Less than a Penny: The Instructional Materials Shortage & How It Shortchanges Students, Teachers, & Schools.* Association of American Publishers Report.

Andrews, Donald W. K. and Moshe Buchinsky. 2001. "Evaluation of a Three-Step Method for Choosing the Number of Bootstrap Repetitions," *Journal of Econometrics*, 103, 345-386

Ballou, Dale. 1996. "Do Public Schools Hire the Best Applicants?" *Quarterly Journal of Economics* 111(1), pp. 97-133.

Black, Dan and Jeffrey Smith. 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence From Matching," *Journal of Econometrics* 121 (2), 99-124.

Borman, Geoffrey D. and N. Maritza Dowling and Carrie Schneck. 2008. "A Multisite Cluster Randomized Field Trial of Open Court Reading," *Education Evaluation and Policy Analysis* 30 (4), 389-407.

Caliendo, Marco and Sabine Kopeinig. 2005. "Some Practical Guidance for the Implementation of Propensity Score Matching," IZA Discussion Paper No. 1588.

Chubb, John and Tom Loveless. 2002. *Bridging the Achievement Gap*, Brookings Institution Press, Washington, D.C.

Fan, Jianqing. 1993. "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196-216.

Finn, Chester. 2004. "The Mad, Mad World of Textbook Adoption" The Thomas B Fordham Foundation and Institute Report. Washington, D.C.

Frölich, Markus. 2004. "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *The Review of Economics and Statistics* 86 (1), 77-90.

Fryer, Roland and Steven Levitt. 2006. "The Black-White Test Score Gap Through Third Grade," *American Law and Economics Review* 8 (2), 249-281.

Ham, John and Xianghong Li and Patricia Reagan. 2006. "Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men," Federal Reserve Bank, Staff Report No. 212.

Heckman, James J. and Jeffrey A Smith. 1995. "Assessing the case for social experiments," *Journal of Economic Perspectives* 9 (2), 85-110.

Heckman, James and Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job-Training Programme," *Review of Economic Studies* 64 (4), 261-294.

Heckman, James and Hidehiko Ichimura and Petra Todd. 1998. "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies* 65 (2), 261-294.

Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics* 86 (1), 4-29.

Jacob, Brian and Lars Lefgren and David Sims. 2008. "The Persistence of Teacher-Induced Learning Gains," NBER Working Paper No. 14065.

Koedel, Cory and Julian R. Betts. 2010. "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation," *Education Finance and Policy* 5 (1), 54-81.

Kullback, Solomon and Richard Leibler. 1951. "On Information and Sufficiency," *Annals of Mathematical Statistics* 22 (1), 79-86.

Lechner, Michael. 2002. "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *The Review of Economics and Statistics* 84 (2), 205-220.

Lechner, Michael. 2004. "Sequential Matching Estimation of Dynamic Causal Models," IZA Discussion Paper No. 1042.

Lechner, Michael and Ruth Miquel (forthcoming). "Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions," *Empirical Economics***.**

Li, Qi and Jeff Racine. 2007. *Nonparametric Econometrics: Theory and Practices*, Princeton University Press, Princeton N.J.

Ludwig, Jens and Douglas Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics* 122 (1) 159-208.

Manski, Charles. 1996. "Learning About Treatment Effects from Experiments with Random Assignment of Treatments," *The Journal of Human Resources* 31 (4), 709-733.

Millimet, Daniel L. and Rusty Tchernis. 2009. "On the Specification of Propensity Scores, with Applications to the Analysis of Trade Policies," *Journal of Business & Economic Statistics* 27 (3), 397-415.

Mueser, Peter R. and Kenneth R. Troske and Alexey Gorislavsky. 2007. "Using State Administrative Data to Measure Program Performance," *The Review of Economics and Statistics* 89 (4), 761-83.

National Research Board. 2004. *On Evaluating Curricular Effectiveness: Judging the quality of K-12 Mathematics Evaluations*, The National Academies Press, Washington DC.

Resendez, Miriam and Mariam Azin. 2007. "The Relationship Between Using Saxon Elementary and Middle-School Math and Student Performance on California Statewide Assessments," Planning Research Evaluation Services.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observtional Studies for Causal Effects," *Biometrika* 70 (1), 41-55.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "The Bias due to Incomplete Matching," *Biometrika* 41 (1), 103-116.

Slavin, Robert E. and Cynthia Lake. 2008. "Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis," *Review of Educational Research*, 78 (3), 427-515.

Smith, Jeffrey and Petra Todd. 2005. "Rejoinder," *Journal of Econometrics* 125 (2), 365-375.

Tulley, Michael. 1989. "The Pros and Cons of State-Level Textbook Adoption," *Publishing Research Quarterly*, 5 (2).

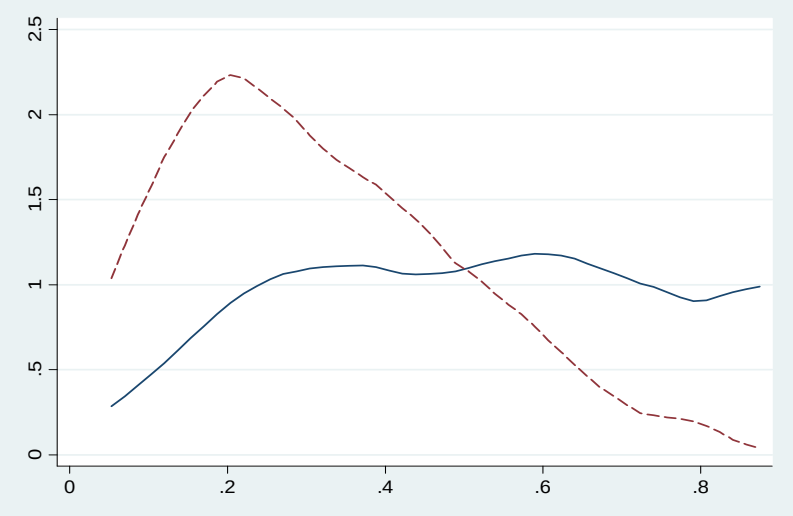
U.S. Department of Health and Human Services, Administration for Children and Families. 2010. *Head Start Impact Study*. Final Report. Washington, DC.

What Works Clearinghouse. 2007. Topic Report: Elementary School Math. Available at: http://ies.ed.gov/ncee/wwc/reports/elementary_math/topic

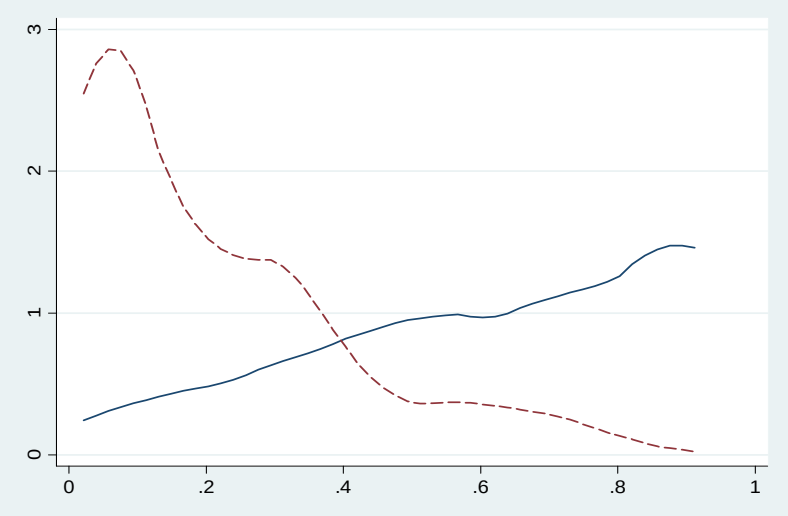
Zhao, Zhong. 2004. "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence" *The Review of Economics and Statistics* 86 (1) 91-107.

Figure 1. Probability Density Functions for Estimated Propensity Scores for Treatment and Control Units on the Common Support in Each Comparison Using 2001 Data (Solid Lines are Treatment Densities, Dashed Lines are Control Densities).

Treatment: Silver-Burdett Ginn (B) Control: Saxon (A)



Treatment: Scott-Foresman (C) Control: Saxon (A)



Treatment: Scott-Foresman (C) Control: Silver-Burdett Ginn (B)

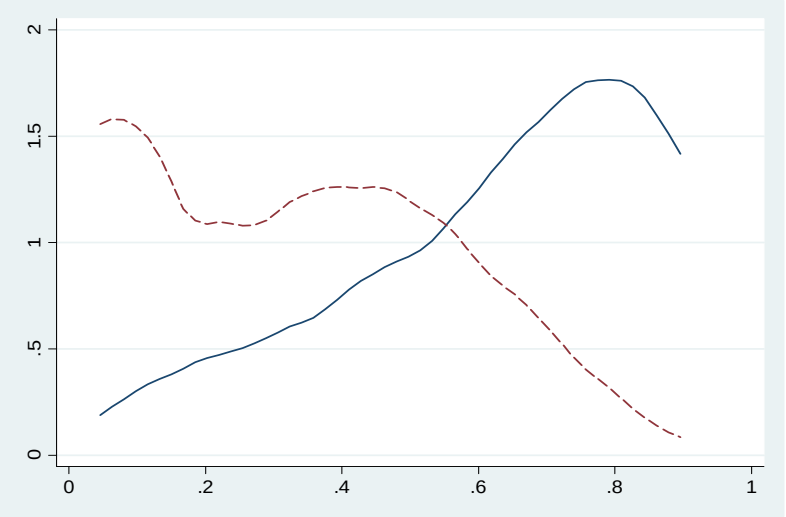


Table 1. Average Characteristics of Schools and Districts, by Adopted Curriculum (1997 values)

	Sample Average	Saxon (A)	Silver (B)	Scott (C)
School-Level Outcomes				
Attendance Rate	96.2	96.3[a]	96.1[a]	96.3
Grade-3 Math Test Score	496.6	496.5	494.2[c]	499.7[c]
Grade-3 Language Test Score	496.7	496.1	495.8	498.7
School-Level Characteristics				
Percent Free Lunch	27.4	24.7[a,b]	28.5[a]	30.5[b]
Percent Reduced Lunch	6.7	7.1[a]	6.3[a]	6.6
Percent Not Fluent in English	1.2	0.7[a]	1.7[a]	1.2
Percent Language Minority	2.6	1.8[a]	3.9[a]	2.6
Percent White	91.3	95.4[a,b]	88.0[a]	88.4[b]
Percent Black	5.6	2.3[a,b]	7.2[a,c]	9.2[b,c]
Percent Asian	0.7	0.4[a,b]	0.9[a]	1.1[b]
Percent Hispanic	2.2	1.8[a,b]	3.7[a,c]	1.1[b,c]
Percent American Indian	0.2	0.1	0.2	0.2
Enrollment (logs)	5.95	5.92	5.97	5.96
N (Schools)	716	311	221	184
District-Level Outcomes				
Attendance Rate	95.8	95.7[b]	95.8	96.1[b]
Grade-3 Math Test Score	498.1	495.8[b]	498.1[a,c]	506.9[b]
Grade-3 Language Test Score	498.9	496.5[a,b]	500.6[a]	505.6[b]
District-Level Characteristics				
Enrollment (logs)	7.72	7.60[a,b]	7.8[a,c]	8.2[b,c]
Total Per-Pupil Revenue (logs)	8.83	8.81[b]	8.84	8.87[b]
Local Per-Pupil Revenue (logs)	7.24	7.18[b]	7.24[c]	7.47[b,c]
Census Information (District Level)				
Median Household Income (logs)	10.81	10.8[a,b]	10.8[a,c]	10.9[b,c]
Share of Population with Low Education	18.2	18.8[b]	19.2[c]	14.3[b,c]
N (Districts)	213	124	56	33

[a] Indicates statistically significant difference at the 10% level between Saxon and Silver-Burdett Ginn adopters.
[b] Indicates statistically significant difference at the 10% level between Saxon and Scott-Foresman adopters.
[c] Indicates statistically significant difference at the 10% level between Silver-Burdett Ginn and Scott-Foresman adopters.
Note: The propensity-score specification also uses italicized information from 1998 – differences in means for these years are not reported for brevity.

Table 2. Balancing details for the 32 covariates included in the multinomial probit specification.

	1992	1993	1994	1995	*1996*	*1999*	2000	2001	2002	2003	2004	2005	2006	2007	2008
Silver (B) to Saxon (A)															
# of unbalanced covariates (p-values below 0.05/0.10)	1/4	0/4	0/3	0/2	*0/2*	*0/2*	0/0	0/0	0/2	0/1	0/0	0/0	0/2	1/2	1/3
Average p-value from balancing tests, all covariates	0.55	0.55	0.55	0.55	*0.55*	*0.55*	0.56	0.56	0.56	0.56	0.57	0.58	0.58	0.57	0.53
Mean Standardized Bias	3.4	2.9	3.8	3.9	*3.3*	*3.5*	3.6	3.3	3.5	3.3	3.0	3.7	3.9	4.2	4.7
Scott (C) to Saxon (A)															
# of unbalanced covariates (p-values below 0.05/0.10)	2/4	4/6	3/6	4/6	*3/5*	*3/6*	3/5	3/5	3/6	5/5	3/5	4/5	5/5	5/5	3/4
Average p-value from balancing tests, all covariates	0.48	0.49	0.49	0.48	*0.50*	*0.49*	0.48	0.48	0.49	0.44	0.45	0.46	0.47	0.47	0.46
Mean Standardized Bias	8.5	5.9	6.1	6.2	*6.1*	*6.0*	6.6	6.3	6.0	7.2	7.6	7.4	7.6	7.6	8.2
Scott (C) to Silver (B)															
# of unbalanced covariates (p-values below 0.05/0.10)	2/5	2/5	2/5	2/5	*1/3*	*0/4*	0/4	1/4	1/4	0/4	0/4	0/4	1/4	3/5	2/4
Average p-value from balancing tests, all covariates	0.48	0.47	0.44	0.46	*0.51*	*0.50*	0.50	0.49	0.50	0.52	0.54	0.54	0.50	0.51	0.54
Mean Standardized Bias	9.6	10.2	8.8	9.3	*9.3*	*9.2*	9.5	9.7	9.8	10.1	10.6	10.6	10.8	10.6	10.8

Note: Columns in italics are for years that are contiguous to the years from which the matching criteria are drawn. Results reported using the samples of treatments and controls that are on the common support in each year for the kernel-matching estimators. The numbers of covariates that fail the balancing tests at the 5 percent level are a subset of those that fail at the 10 percent level.

Table 3. Kullback-Leibler (KL) Information Criteria by Curriculum Comparison.

Comparison	KL Information Criterion
B and A	0.63
C and A	1.58
C and B	0.91

Note: Based on 2001 sample of schools.

Table 4. Estimates of Math Curricular Effectiveness on Math Test Scores for Partially and Fully-Exposed Cohorts. All Comparisons.

	1999	2000	**2001**	**2002**	**2003**	**2004**	2005	2006
Treatment: B Control: A								
OLS	0.124	0.162	**0.354**	**0.356**	**0.374**	**0.268**	0.293	0.250
	(0.105)	(0.101)	(**0.095**)**	(**0.087**)**	(**0.099**)**	(**0.131**)*	(0.104)**	(0.110)*
Kernel Matching	0.144	0.191	**0.396**	**0.400**	**0.401**	**0.279**	0.318	0.253
	(0.139)	(0.145)	(**0.125**)**	(**0.102**)**	(**0 .116**)**	(**0.135**)*	(0.130)**	(0.132)†
LLR Matching	0.154	0.173	**0.397**	**0.398**	**0.398**	**0.273**	0.308	0.259
	(0.184)	(0.153)	(**0.117**)**	(**0.122**)**	(**0.126**)**	(**0.147**)†	(0.138)*	(0.134)†
Treatment: C Control: A								
OLS	0.130	-0.013	**0.187**	**0.261**	**0.208**	**0.014**	0.109	0.183
	(0.120**)**	(0.134**)**	(**0.104)†**	(**0.096)****	(**0.110)†**	(**0.119)**	(0.104**)**	(0.119**)**
Kernel Matching	0.117	0.010	**0.215**	**0.270**	**0.272**	**-0.042**	0.113	0.150
	(0.169)	(0.184)	(**0.158**)	(**0.122**)*	(**0.124**)*	(**0.118**)	(0.187)	(0.187)
LLR Matching	0.128	0.135	**0.169**	**0.295**	**0.301**	**0.032**	0.085	0.141
	(0.248)	(0.269)	(**0.220**)	(**0.156**)†	(**0.199**)	(**0.224**)	(0.243)	(0.354)
Treatment: C Control: B								
OLS	0.008	-0.160	**-0.100**	**-0.186**	**-0.285**	**-0.271**	-0.181	-0.083
	(0.100)	(0.123)	(**0.117)**	(**0.129)**	(**0.166)**†	(**0.162)**†	(0.129**)**	(0.139**)**
Kernel Matching	-0.088	-0.237	**-0.165**	**-0.164**	**-0.331**	**-0.275**	-0.208	-0.148
	(0.255)	(0.274)	(**0.230**)	(**0.183**)	(**0.193**)†	(**0.204**)	(0.239)	(0.249)
LLR Matching	-0.072	-0.230	**-0.149**	**-0.122**	**-0.302**	**-0.236**	-0.163	-0.163
	(0.657)	(0.531)	**(0.652)**	**(0.898)**	**(0.358)**	**(0.219)**	(0.484)	(0.798)
N(A)	309	307	**307**	**305**	**300**	**294**	286	287
N(B)	220	219	**219**	**213**	**213**	**212**	210	207
N(C)	184	182	**182**	**181**	**176**	**174**	169	163

Notes: Bolded columns are for the fully-exposed cohorts. Matching estimators impose the common support restriction. Standard errors in parentheses are clustered at the district level for all estimates, and bootstrapped using 250 repetitions for the matching estimators.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 5. Average 2004 Curriculum Adoptions in Math by District for the Four Most Common Curricula from the 2004 Adoption Cycle.

	1998 Uniform Math Adoptions – Grades 1 Through 3				
	All	Saxon (A)	Silver-Burdett Ginn (B)	Scott-Foresman (C)	Other
2004 Math Adoptions					
Grade 1					
Saxon	0.48	0.76	0.25	0.12	0.21
Harcourt	0.19	0.07	0.32	0.35	0.24
Houghton Mifflin	0.10	0.06	0.11	0.21	0.15
Scott-Foresman	0.09	0.07	0.07	0.15	0.18
Grade 2					
Saxon	0.48	0.77	0.25	0.09	0.24
Harcourt	0.19	0.08	0.32	0.35	0.21
Houghton Mifflin	0.10	0.06	0.11	0.21	0.15
Scott-Foresman	0.09	0.05	0.07	0.18	0.18
Grade 3					
Saxon	0.48	0.76	0.23	0.09	0.24
Harcourt	0.18	0.08	0.32	0.35	0.21
Houghton Mifflin	0.12	0.07	0.14	0.21	0.15
Scott-Foresman	0.09	0.06	0.05	0.21	0.15
Grade 4					
Saxon	0.47	0.73	0.21	0.12	0.24
Harcourt	0.18	0.09	0.30	0.35	0.21
Houghton Mifflin	0.12	0.09	0.12	0.18	0.15
Scott-Foresman	0.11	0.07	0.11	0.21	0.15
Grade 5					
Saxon	0.47	0.74	0.21	0.18	0.22
Harcourt	0.18	0.08	0.30	0.32	0.22
Houghton Mifflin	0.10	0.07	0.11	0.18	0.16
Scott-Foresman	0.11	0.08	0.11	0.18	0.16
N	286	128	57	34	33

Notes: N indicates the number districts where we observe a 2004 math curriculum adoption and at least one grade-3 math test score between 1998 and 2008. The "other" category includes all districts that did not adopt any of the "big three" curricula in any grade during the 1998 adoption cycle. Districts that adopted at least one of the big-three curricula *non-uniformly* during the 1998 adoption cycle are included only in the "all" category.

Table 6. Falsification Estimates of Math Curricular Effectiveness, Estimated Using Math Test Scores for Grade-3 Cohorts Who Were Never Exposed to the Curricula of Interest. All Comparisons.

	1992	1993	1994	1995	1996	2007	2008
Treatment: B Control: A							
Kernel Matching	-0.120	0.072	-0.019	0.079	0.094	0.091	0.192
	(0.112)	(0.135)	(0.120)	(0.137)	(0.129)	(0.117)	(0.127)
Treatment: C Control: A							
Kernel Matching	-0.326	-0.046	-0.011	-0.035	-0.045	-0.020	-0.050
	(0.162)*	(0.174)	(0.146)	(0.186)	(0.153)	(0.157)	(0.270)
Treatment: C Control: B							
Kernel Matching	-0.171	0.077	0.032	0.072	-0.066	-0.147	-0.235
	(0.274)	(0.277)	(0.237)	(0.294)	(0.280)	(0.202)	(0.263)
N(A)	301	304	304	306	308	284	280
N(B)	209	210	213	216	220	205	201
N(C)	179	179	182	182	182	163	162

Notes: Matching estimators impose the common support restriction. Standard errors in parentheses are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 7. Falsification Estimates of Math Curricular Effectiveness, Estimated Using Math Test Scores for Grade-6 Cohorts who were Never Exposed to the Curricula of Interest. Comparison of *B* and *A* only.

	1992	1993	1994	1995	1996	1999	2000	2001
Treatment: B Control: A								
Kernel Matching	-0.126	-0.290	-0.055	-0.133	0.045	0.016	-0.190	-0.100
	(0.155)	(0.165)†	(0.158)	(0.139)	(0.142)	(0.177)	(0.151)	(0.130)
N(A)	205	208	213	213	218	212	205	204
N(B)	117	118	122	125	127	122	120	120

Notes: Matching estimators impose the common support restriction. Standard errors in parentheses are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 8. Estimates of Math Curricular Effectiveness, Estimated Using Reading Test Scores for all Grade-3 Cohorts. All Comparisons.

	1992	1993	1994	1995	1996	1999	2000	**2001**	**2002**	**2003**	**2004**	2005	2006	2007	2008
Treatment: B Control: A															
Kernel Matching	-0.152	-0.036	-0.078	0.082	0.135	0.160	0.186	**0.197**	**0.228**	**0.150**	**0.084**	0.027	0.044	-0.084	0.069
	(0.110)	(0.131)	(0.130)	(0.129)	(0.141)	(0.141)	(0.146)	(**0.146**)	(**0.123**)†	(**0.120**)	(**0.127**)	(0.148)	(0.122)	(0.118)	(0.117)
Treatment: C Control: A															
Kernel Matching	-0.200	-0.126	-0.107	-0.154	-0.161	0.023	0.048	**0.036**	**-0.043**	**0.037**	**-0.030**	-0.080	0.184	0.028	0.084
	(0.156)	(0.175)	(0.178)	(0.206)	(0.178)	(0.207)	(0.237)	(**0.221**)	(**0.159**)	(**0.177**)	(**0.205**)	(0.205)	(0.208)	(0.205)	(0.212)
Treatment: C Control: B															
Kernel Matching	-0.023	0.149	0.118	0.009	-0.179	-0.172	-0.143	**-0.166**	**-0.222**	**-0.125**	**-0.095**	-0.020	0.113	0.065	-0.014
	(0.294)	(0.305)	(0.268)	(0.288)	(0.290)	(0.281)	(0.321)	(**0.289**)	(**0.245**)	(**0.259**)	(**0.213**)	(0.297)	(0.297)	(0.262)	(0.303)
N(A)	301	304	304	306	308	309	307	**307**	**305**	**300**	**294**	286	287	284	280
N(B)	209	210	213	216	220	220	219	**219**	**213**	**213**	**212**	210	207	205	201
N(C)	179	179	182	182	182	184	182	**182**	**181**	**176**	**174**	169	163	163	162

Notes: Bolded columns are for the fully-exposed cohorts. Matching estimators impose the common support restriction. Standard errors in parentheses are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Table 9. Persistence Effects. Estimated Curriculum Effects for Grade-6 Cohorts who were Partially or Fully Exposed. Comparison of *B* and *A* only.

	2002	2003	**2004**	**2005**	**2006**	**2007**	2008
Treatment: B Control: A							
Kernel Matching	-0.064	0.141	**0.156**	**0.077**	**0.007**	**-0.023**	-0.016
	(0.151)	(0.146)	**(0.199)**	**(0.173)**	**(0.150)**	**(0.169)**	(0.159)
N(A)	200	189	**174**	**165**	**163**	**160**	156
N(B)	118	115	**105**	**101**	**97**	**94**	93

Notes: Bolded columns are for the fully-exposed cohorts. Matching estimators impose the common support restriction. Standard errors in parentheses are clustered at the district level and bootstrapped using 250 repetitions.

** Denotes statistical significance at the 1 percent level or better

* Denotes statistical significance at the 5 percent level or better

† Denotes statistical significance at the 10 percent level or better

Appendix A
Supplementary Tables

Appendix Table A.1. Data Sample Details.

	Schools	% of Universe	Districts	% of Universe
Universe*	1115		294	
Missing Information:				
District-reported curriculum adoption	3	0.3	3	1.0
District outcome variables (1997)	2	0.2	2	0.7
School outcome variables (1997)	23	2.2	1	0.3
District finance or enrollment data (1997, 1998)	2	0.2	1	0.3
School enrollment or demographic data (1997, 1998)	82	7.3	12	4.0
Did not use one of the primary curricula in grades one, two or three	211	18.9	38	12.9
Used only primary curricula, but did not uniformly adopt	76	6.8	24	8.2
Final Sample	*716*	*64.2*	*213*	*72.4*

* The universe consist of those schools and districts for which any information was reported in 1997, and at least one grade-3 math test score was reported for an exposed cohort (1999-2006).

Appendix Table A.2. Scaling Factors Used to Convert Estimation Metric from School-Level Distribution to Individual-Level Distribution for Grade-3 Math Scores.

Year	Standard Deviation of Distribution of School Scores	Standard Deviation of Distribution of Individual Scores	Approximate Scaling Factor
1992	2.8	N/A	N/A
1993	2.9	N/A	N/A
1994	2.8	N/A	N/A
1995	2.8	N/A	N/A
1996	1.9	N/A	N/A
1999	21.3	N/A	N/A
2000	20.5	61.0	0.34
2001	21.0	61.4	0.34
2002	19.9	59.7	0.33
2003	20.7	60.9	0.34
2004	22.5	63.1	0.36
2005	21.0	62.2	0.34
2006	20.0	64.3	0.31
2007	21.3	65.4	0.33
2008	22.5	63.7	0.35

Appendix B
Bandwidth Selection

We use standard leave-one-out cross validation (C-V) to obtain fixed bandwidths for the kernel and LLR matching estimators. The grid search for kernel and LLR matching is over the range (0.005, 2.0). Using Frölich's (2004) notation, the C-V approach selects the optimal bandwidth, h_{CV}, by solving the following minimization problem for control observations:[38]

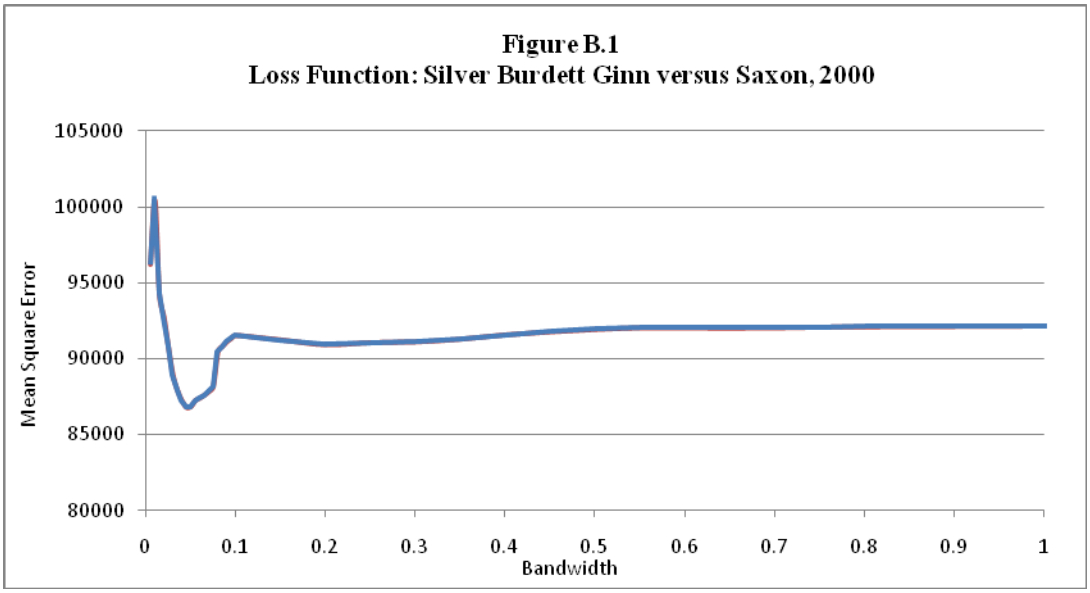
$$h_{CV} = \arg\min(h)\sum_{q=1}^{Q}(Y_q - \hat{m}_{-q}(p_q))^2$$

where q indexes the sample of control units, Y is the outcome (test score) and $\hat{m}_{-q}(p_q)$ is the estimate of the mean outcome among the control observations, excluding observation q, conditional on the estimated propensity score for unit q.

As has been reported in other contexts (see, for example, Ludwig and Miller, 2007), the loss function used to select the bandwidth is fairly flat in most of our comparisons. Therefore, we use a combination of conventional C-V and "visual inspection" to identify the appropriate bandwidth for each of our matching estimators.

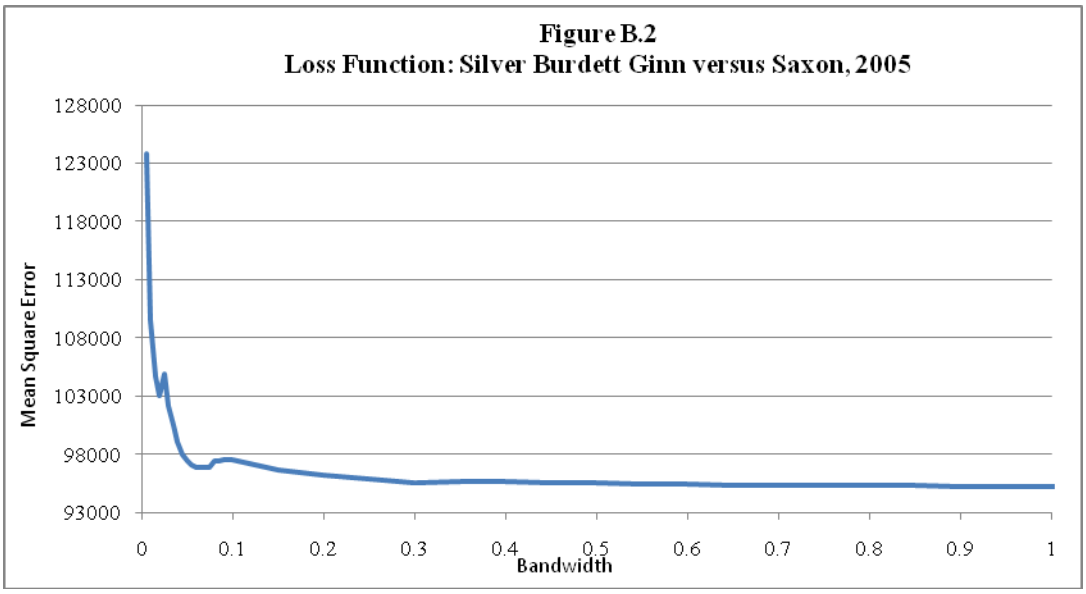
First, Figure B.1 illustrates a case where cross-validation produces a clear bandwidth choice at the global minimum of the loss function, for our comparison between *B* and *A* in 2000 using the kernel matching estimator. In this case we use bandwidth at the global minimum, 0.048.

Figure B.1
Loss Function: Silver Burdett Ginn versus Saxon, 2000



[38] In our case the definition of "treatment" and "control" in arbitrary and therefore, we could use either group. We use the largest group in each comparison.

Next, Figure B.2 illustrates a case where cross-validation suggests an optimal bandwidth at the edge of our grid search, for our comparison between *B* and *A* in 2005 using the kernel matching estimator. For this comparison we use a bandwidth of 0.062, which occurs just prior to the narrowly upward sloping portion of the curve.



We describe our bandwidth selection procedure for the comparison in Figure B.2 as a combination of cross-validation and visual inspection. Because the flat region of the curve has a mild negative slope, the mechanical application of the C-V procedure would produce a bandwidth at the edge of our grid search, 2.0. However, by visual inspection, we can see that there is very little difference in the loss function between the bandwidth determined mechanically by the C-V procedure and a much narrower bandwidth selected after the initial drop in the loss function. We ultimately use the narrower bandwidth in this and similar cases because the efficiency gains associated with the wider bandwidth will be minimal, and the narrower bandwidth should reduce bias in the estimates.

Across our grade-3 comparisons spanning the entire data panel, our approach of combining C-V with visual inspection yields a bandwidth at the global minimum of the loss function 40 percent of the time. In the remaining cases where the global minimum occurs at the edge of our grid search, the average increase in the loss function that we observe by choosing an interior bandwidth is 1.3 percent, with a maximum increase of 2.9 percent in one instance. Details about our bandwidth selection process for each estimator in the paper are available upon request.

Finally, that cross validation produces large flat regions in the loss function in most of our comparisons provides some indirect evidence that curriculum adoptions are not meaningfully correlated with other, unobservable determinants of school performance. The flat regions suggest that as increasingly non-comparable units (as measured by the propensity score) are used as comparisons for one another, there is minimal change in their measured outcomes. Such conditions will certainly be favorable to a non-experimental analysis.